Characterizing Home Network Traffic: An Inside View

Kuai Xu[1], Feng Wang[1], Lin Gu[2], Jianhua Gao[3], and Yaohui Jin[4,5]

[1] Arizona State University
[2] Hong Kong University of Science and Technology
[3] Wuhan University
[4] Shanghai Jiaotong University
[5] State Key Laboratory of Advanced Optical Communication Systems and Networks

Abstract. The rapid spread of residential broadband connections and Internet-capable consumer devices in home networks has changed the landscape of Internet traffic. To gain a deep understanding of Internet traffic for home networks, this paper develops a traffic monitoring platform that collects and analyzes home network traffic via programmable home routers and traffic profiling servers. Using traffic data captured from real home networks, we present traffic characteristics in home networks, and then apply principal component analysis to uncover temporal correlations among application ports. To the best of our knowledge, this paper is the first study to characterize network traffic of Internet-capable devices from inside home networks.

1 Introduction

In recent years, the rapid growth of Internet-capable devices in the home and residential broadband access has driven the rising adoptions of home networks. The availability of home networks not only creates new application opportunities such as remote health care and Internet television, but also changes the distribution of Internet traffic, e.g., a recent study shows that video streaming via Netflix accounts for 32.7% of peak downstream traffic in United States [1]. As home networks become an important part of the Internet ecosystem, it is very crucial to understand network traffic between the Internet and home networks as well as the traffic exchanged within home devices.

Most home users lack technical expertise to manage the increasingly complicated home networks, and an extensive body of research have focused on how to simplify network management tasks for home users [2–6]. Several recent studies have been devoted to understanding traffic characteristics of home networks using aggregated and sampled traffic collected from edge routers in Internet service providers [7–9]. However, these measurement studies stand from the perspective of outside home networks, thus lack the visibility of *what is happening in home networks*. The in-depth understanding of home network traffic could aid home users in effectively securing and managing home networks.

X. Wang et al. (Eds.): WASA 2012, LNCS 7405, pp. 60–71, 2012.

In this paper we focus on understanding traffic characteristics in home networks. Towards this end, we first develop a traffic monitoring platform that collects network flow streams via traffic profiling servers and programmable home routers that connect home networks and the Internet via home gateways such as DSL or cable modems. Using traffic data collected from real home networks, we analyze traffic patterns of connected devices in home networks, and characterize the volume, behavior and temporal features of home network traffic.

Our findings on temporal characteristics of application ports lead us to explore principal component analysis (PCA) to uncover temporal correlations among these ports. The experiment results show that there indeed exist several application port clusters in home networks with each cluster exhibiting distinct traffic patterns. For example, one cluster consists of major canonical applications including 80/TCP (HTTP), 443/TCP (HTTPS), 53/UDP (DNS), while another cluster contains a group of unknown ports with all traffic sent to temporary servers running on Amazon Elastic Compute Cloud (Amazon EC2). Closer examinations reveal that all the traffic in the latter cluster are associated with suspicious activities.

The contributions of this paper are two-fold. First, we develop a traffic monitoring platform that automatically collects, analyzes and makes sense of network traffic for Internet-capable devices in home networks. Secondly, we present traffic characteristics of home networks, and apply principal component analysis to uncover temporal correlations among application ports. To the best of our knowledge, this paper is the first study to characterize traffic patterns of Internet-capable home devices from the inside perspective.

The remainder of this paper is organized as follows. Section 2 describes the traffic monitoring platform we developed for home networks, while Section 3 presents the basic characteristics of home network traffic, and applies principal component analysis to uncover temporal correlations of application ports in home network traffic. Section 4 discusses related work, and Section 5 concludes this paper and outlines the future work.

2 Traffic Monitoring Platform for Home Networks

To understand what is happening in home networks, we develop a real-time behavior monitoring platform to collect and analyze network traffic for Internet-capable devices in the home [10]. The monitoring platform captures network traffic via programmable home routers, which connect home networks with the Internet through home gateways such as cable or DSL modems. Using a Linux distribution for embedded devices, OpenWrt [11], we configure a programmable home router and export network flows traversing through all the interfaces of the router to a traffic profiling server running in the same home network. The continuous network flows, aggregated from IP packets, contain a number of important features for our traffic analysis including the start and end time-stamps, source IP address (`srcIP`), destination IP address (`dstIP`), source port number (`srcPort`), destination port number (`dstPort`), and protocol, packets and bytes.

Many host-based monitoring systems are also able to collect these traffic flows on individual devices, e.g., Windows and Linux machines, however such host-based approaches are very difficult to deploy across all the possible devices due to the high heterogeneity of Internet-capable devices in the home.

Compared with incoming and outgoing traffic of home networks, the overhead of transferring flow data from programmable home routers to traffic profiling servers is not significant. Figure 1 shows the overhead of collecting traffic data from programmable home routers (top figure), the bandwidth usages of outgoing traffic (middle figure) and incoming traffic (bottom figure) of one home network that deploys the platform. As shown in the top figure, the network flow data exported by home routers consumes less than 4Kbps bandwidth, which is much smaller than outgoing and incoming traffic illustrated in the middle and bottom graphs. In general, the network bandwidth usage of incoming traffic towards home networks is larger than that of outgoing traffic, as most of Internet activities in these home networks are Web browsing, email communications, and video streaming.

The availability of the traffic monitoring platform makes it possible for us to analyze data traffic exchanged between home devices and Internet end hosts, as well as data traffic exchanged among home network devices. Making sense of these traffic could not only assist home users in understanding *what is happening in home networks*, but also help detect anomalous traffic towards home networks or originating from compromised home devices. In the next section, we will use traffic data collected from real home networks that deploy the traffic monitoring platform to characterize network traffic of Internet-capable home devices from a

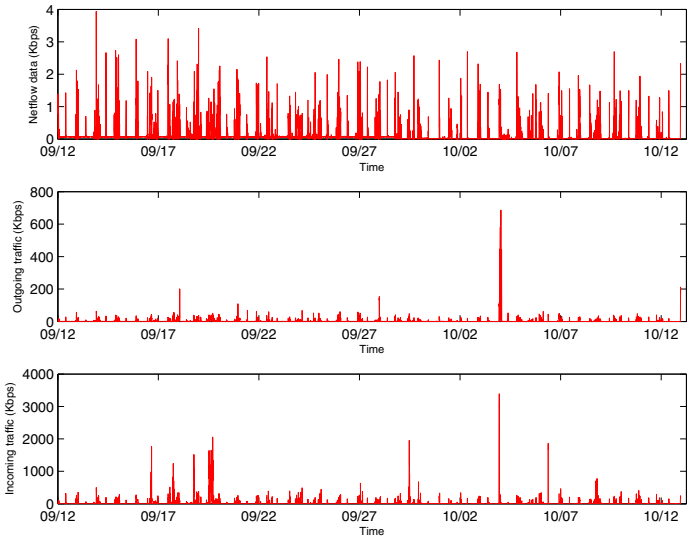


Fig. 1. Bandwidth usage of data collection (top figure), outgoing traffic (middle figure) and incoming traffic (bottom figure) of home networks

variety of traffic information including *volume features* measured by the numbers of flows, packets and bytes, *social features* through analyzing IP addresses and application ports, and *temporal dynamics* of these traffic. Each of these traffic features captures the behavior of home devices from a unique perspective. Combined together, they provide a broad picture of home network traffic, and more importantly, reveal interesting traffic activities in home networks.

3 Characterizing Home Network Traffic

In this section, we first describe data-sets used in this study and present the general characteristics of home network traffic. Subsequently, we explore principal component analysis to analyze temporal correlations among application ports for uncovering clusters of application ports sharing significant temporal patterns in network traffic.

3.1 Datasets

The traffic data used in this study is collected from two home networks (home network A and home network B) that deploy our traffic monitoring platform during one-month time span from `09/12/2011` to `10/12/2011`. The numbers of total devices in home networks A and B are 6 and 3, respectively. Figure 2 shows the number of *online* devices in home network A over time. As illustrated in Figure 2, the number of *online* devices in home network A observed during 5-min time bins varies from 0 to 6, reflecting Internet usage patterns of these devices during this one-month time period. Note that the number of home devices remaining above 1 between 09/12 and 09/28 is due to a probing program continuously running on one home device to measure end-to-end performance to

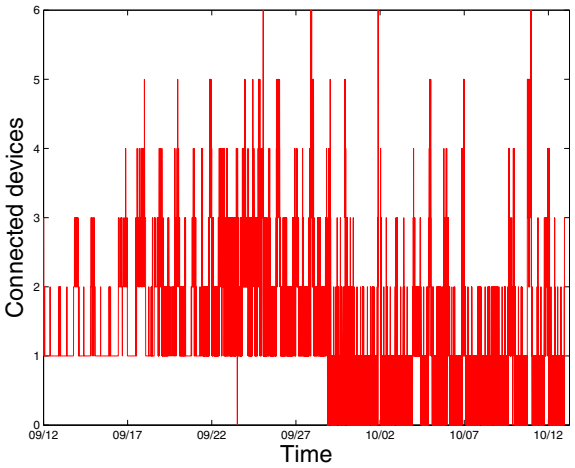


Fig. 2. The number of *online* devices in the home network A over time

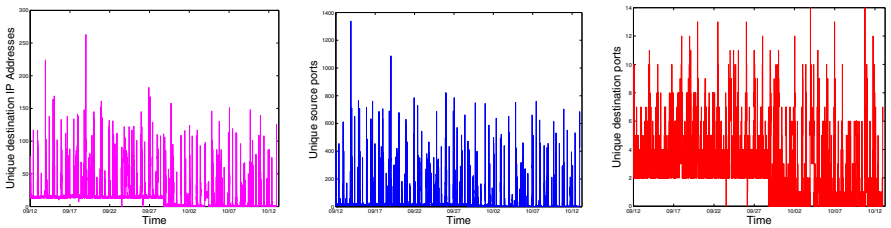
a number of distributed servers. These devices collectively have communicated with over 4,800 unique end hosts on the Internet from 529 different autonomous systems (ASes) during this period. Similarly, the devices in home network B collectively communicate with over 4,400 end hosts from 726 ASes.

3.2 Traffic Characteristics

We study the traffic characteristics of home networks by firstly examining IP addresses and application ports over time, since they reflect *whom do home devices communicate with* and *what applications do home devices use*. Figures 3[a-c] illustrate the numbers of unique destination IP addresses, unique source ports and unique destination ports for the outgoing traffic during 5-min time bins over time, respectively.

Our first interesting observation lies in the large number of unique destination IP addresses during 5-min time bins, as shown in Figure 3[a]. Closer examination revealed that a single visit to a major content-rich Web portal could trigger tens of TCP connections to different Web servers, and the large number of destination IP addresses actually correspond to legitimate Web servers visited by home users. For example, our empirical experiment of visiting the front page of `www.cnn.com` with a Firefox browser finds that loading the entire page requires the browser to talk with 18 different IP addresses from a variety of Internet service and content providers including `Facebook` (social network site), `Google` (search engine), `Limelight Networks` (content deliver network), `Rackspace Hosting` (cloud service provider), `Valueclick` (online advertising), and `cnn` itself.

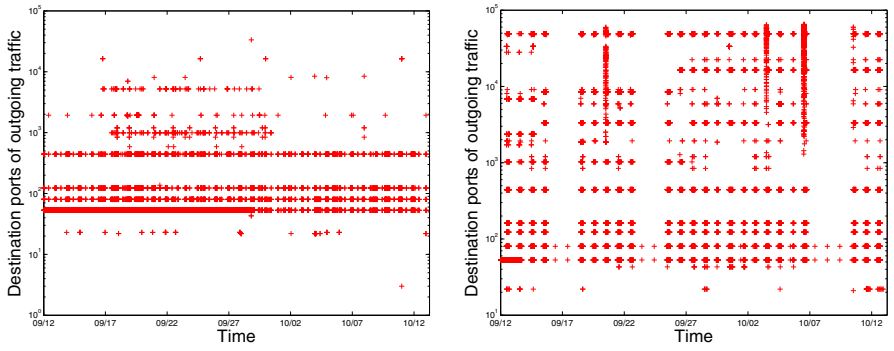
The second observation from Figure 3 is that the number of unique destination ports for outgoing traffic in home networks is far less than that of unique source ports. The small number of destination ports in outgoing traffic provides a simple and natural classification on home network traffic, thus we follow a port-driven approach for further traffic analysis. Specifically, we separate outgoing traffic flows into distinct groups based on their destination ports in order to gain an in-depth understanding on network traffic of each individual destination port. Similarly, we group incoming traffic flows into distinct groups based on their source destination ports.



(a) dest. IP addresses (b) source ports (c) destination ports

Fig. 3. The number of unique IP addresses and ports in outgoing traffic for home network A over time

Figures 4[a][b] illustrate the temporal frequency of all destination ports for home network A and B during one-month time period, respectively. It is interesting to find three types of temporal patterns among these ports. The first type of destination ports are consistently observed during all days. For example, port 80/TCP is observed in all days during the one-month period in both networks. The second type of ports are observed during several days, while the last type includes ports that are only observed in one or two days suggesting these infrequent ports might be associated with unusual or anomalous traffic. Similar observations hold for the source ports in the incoming traffic towards home networks. More interestingly, Figures 4[a][b] also reveal temporal correlations among groups of applications ports that consistently show up around approximately the same times. This observation motivates us to explore correlation analysis techniques to understand the reasons behind such temporal correlations.



(a) home network A (b) home network B

Fig. 4. Time-series observations of destination ports in outgoing traffic

3.3 Temporal Correlation Analysis of Application Ports

To explore temporal correlation among application ports in home networks, we propose to use principal component analysis (PCA) to analyze traffic patterns of network applications. PCA is a widely used technique in network traffic analysis [12, 13] due to its ability of analyzing multivariate data and locating interrelated variables [14].

Let p and t denote the total number of ports observed in the data and the total number of time bins. Our initial step is to construct a $p \times t$ matrix X, where $x_{i,j}$ denotes the total number of network flows for the destination port i ($i = 1, 2, \ldots, p$) in the outgoing traffic (or the source port i in the incoming traffic) during the j-th ($j = 1, 2, \ldots, t$) time period. The vector ${x_i}^T$ reflects a time-series of observations for the application port i. Next we obtain the covariance matrix S, p non-decreasing ordered eigenvalues, $\lambda_1, \lambda_2, \ldots, \lambda_p$, and the corresponding

eigenvectors $\alpha_1, \alpha_2, \ldots, \alpha_p$, where where s_{ab} is the covariance of two application ports a and b, and $S\alpha_i = \lambda_i \alpha_i$, for $1 \leq i \leq p$.

The p principal components of the matrix X can be derived by projecting the matrix onto the p eigenvectors, i.e., $PC_i = \alpha_i^T X$, $i = 1, 2, \ldots, p$. As $var(PC_i) = var(\alpha_i^T X) = \alpha_i^T X \cdot X^T \alpha_i = \alpha_i^T S \alpha_i = \lambda \alpha_i^T \alpha_i = \lambda_i$, the variance captured by the i-th principal component is essentially the i-th eigenvalue λ_i.

PCA transforms the space of the p observed variables in the original matrix X into a new space of p principal components $\{PC_i\}$, $i = 1, 2, \ldots, p$. Figure 5 shows the distribution of the eigenvalues using the matrix constructed with the one-month traffic data from home network A. As shown in Figure 5, a few largest eigenvalues account for the majority of the variance in the original matrix, suggesting that the corresponding top principal components capture most variances.

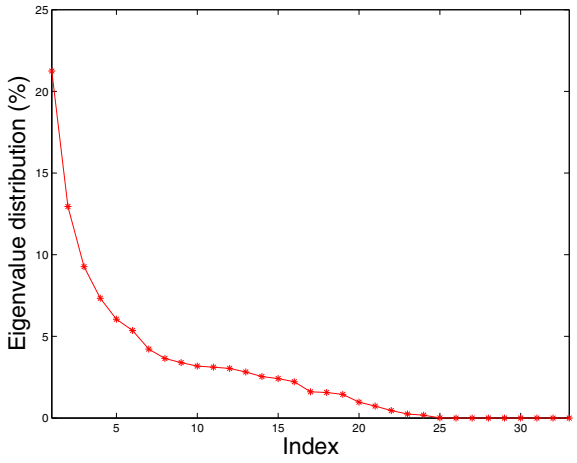


Fig. 5. Eigenvalue distribution of the matrix constructed with the one-month traffic data for home network A

Thus, the final step of the PCA process is to project the original data-set onto a subspace with a smaller dimensionality to get approximate representations while retaining the majority of the variance in the original data-set. Specifically, we require that the largest m eigenvalues that are larger than a fixed threshold such that each selected principal component captures a non-trivial variance in the original data-sets. In the experiment, we use 5% of the total variances as the threshold for determining the value of m.

The principal component PC_i can also be represented as: $PC_i = \alpha_i^T X = [\alpha_{i1} x_1 + \ldots + \alpha_{ip} x_p]^T = [\sum_{j=1}^{p} \alpha_{ij} x_j]^T$, where α_{ij}, $j = 1, \ldots, p$, is the coefficient of x_j for PC_i. The coefficient value α_{ij} reflects the contribution or influence of the application port j to the variance obtained by the i-th component. Such relationship between principal components and observed variables leads to the discovery of a cluster of application ports that contribute similar influence towards

the same principal components because of the inherent temporal correlations among these ports. As a result, we group the application ports that contribute similar high influence towards the variance of each of the top principal components into a distinct `srcPort` cluster for incoming traffic (or a `dstPort` cluster for outgoing traffic). In other words, PCA discovers the clusters of application ports that exhibit significant correlations in the temporal traffic patterns.

Table 1 lists the membership of the 6 `dstPort` clusters discovered via the principal component analysis using one-month traffic data collected in home network *A*. $Cluster_1$ includes port 43/TCP and consecutive ports 33435-33440/UDP. The in-depth analysis shows that the flows associated with 43/TCP are legitimate *whois* traffic towards *Team Cymru IP to AS mapping service*, while all traffic associated with ports 33435-33440/UDP were sent towards an unknown server and failed to get response from the server. The legitimate traffic on port 43/TCP and suspicious traffic on 33435-33440/UDP were observed during the same time window, which explain these seven ports to be grouped as a single `dstPort` cluster. Although $Cluster_1$ includes a service port 43/TCP, the majority of ports, 33435-33440/UDP, does reflect anomalous traffic activity from one home network device. $Cluster_2$ includes four canonical ports (i.e., DNS, HTTP, HTTPS, and NTP), which are used by home network devices on a daily basis and thus naturally form a `dstPort` cluster.

$Cluster_3$ includes three consecutive ports 16384-16386/UDP, which was sent by the FaceTime video calling application on an iPhone device. This cluster indicates that many user-installed applications or vendor-installed applications could use non-traditional ports for data communications with end hosts on the Internet. Such practices make it more challenging to differentiate anomalous or legitimate traffic on unusual ports. $Cluster_4$ includes three ports, i.e., 843/TCP, 1200-1201/TCP. Closer examinations reveal that a Windows laptop communicated with seven different instances in Amazon EC2 Cloud on these three ports *simultaneously* during 9 different days over the first two weeks. As home users are not aware of any application involving these ports and servers, these traffic is likely sent by a malware on the compromised laptop. $Cluster_5$ includes two ports 1863/TCP and 7001/UDP used by Windows MSN messenger, while $Cluster_6$ includes two ports 993/TCP and 5223/TCP, which are used by GMail and Apple Push Notification service running on the iPhone device that connects to the home network over Wi-Fi.

These experiment results with real home network traffic confirm that there indeed exist a variety of `dstPort` clusters that group applications ports with strong temporal correlations. Some of these clusters, e.g., $Cluster_1$ and $Cluster_4$ in Table 1, even lead to surprising findings on suspicious network traffic originating from home network devices that might be compromised by Internet malwares. Therefore, characterizing network traffic for Internet-capable devices in the home could not only provide valuable insight on behavior patterns of these connected devices, but also help improve the security and management of home networks.

Table 1. `dstPort` clusters discovered via PCA on temporal correlation

`dstPort` Cluster	Port Number	Application	User-aware
1	43/TCP	whois	Yes
	33435/UDP	unknown	No
	33436/UDP	unknown	No
	33437/UDP	unknown	No
	33438/UDP	unknown	No
	33439/UDP	unknown	No
	33440/UDP	unknown	No
2	53/UDP	DNS	Yes
	80/TCP	Web/HTTP	Yes
	123/UDP	NTP	Yes
	443/TCP	Web/HTTPS	Yes
3	16384/UDP	FaceTime	Yes
	16384/UDP	FaceTime	Yes
	16386/UDP	FaceTime	Yes
4	843/TCP	unknown	No
	1200/TCP	unknown	No
	1201/TCP	unknown	No
5	1863/TCP	MSN	Yes
	7001/UDP	MSN	Yes
6	993/TCP	IMAP over SSL	Yes
	5223/TCP	AppPush Notification Service	Yes

4 Related Work

Unlike enterprise networks which have dedicated network professionals to manage and operate the networks, securing home networks has been a considerable challenge, as most home users do not have sufficient technical expertise and knowledge to manage and secure the networks [15]. As a result, connected devices in home networks are targets and victims of virus, worms, and botnets, and become a major source of spams and a part of botnets. In [16], Feamster proposes to outsource the management and operations of home networks to a third party that has expertise of network operations and security management. In [17], Yang et al. study network management tools that are currently deployed in home networks via interviewing 25 home networks users, and report user experiences of these network management tools. To aid in troubleshooting and managing home networks, [18] proposes to build a home network data recorder system as a general-purpose logging platform to record what is happening in home networks. Many researches have also focused human computer interactions in home networks [2–4], troubleshooting and diagnosis [5, 6], and broadband network sharing among different Internet service providers [19].

Home network performance has recently drawn significant attentions from the research community. A recent work [20] performs controlled experiments in a lab environment for evaluating the impact of home networks on end-to-end performance of end systems. In addition, several commercial or open source tools have been developed for measuring and diagnosing Internet properties of end users. For example, Netalyze [21], a network measurement and diagnosis service, tests a wide variety of functionalities at network, transport and application layers

for end users' Internet connectivity in edge networks such as home networks. Kermit, a network probing tool, was developed in [22] to visualize the broadband speed and bandwidth usage for home users. [23] measures and analyzes the behavior characteristics of a variety of home gateways such as DSL and cable modems, including NAT binding timeout, throughput, and protocol support, and their influence on network performance and user experience. A recent work [24] measures network access link performance directly from home gateway devices, and has inspired us to characterize network traffic from inside home networks through programmable home routers.

As residential broadband users continue to grow, many studies have been devoted to measure and characterize residential broadband networks [7–9]. However, all of these studies stand from outside home networks, and lack the visibility of the home networks, such as home network architecture, diversity of end hosts. For example, [7] examines the growth of residential user-to-user traffic in Japan, a country with a high penetration rate of residential broadband access, and studies the impact of these traffic on usage patterns and traffic engineering of commercial backbone networks. In addition, [8] studies several properties of broadband networks, including link capacities, round-trip times, jitter, and packet loss rates using active TCP and ICMP probes, while [9] passively collects packet-level traffic data of residential networks at aggregated routers of a large Internet service provider, and analyzes dominant characteristics of residential traffic including network and transport-level features, prominent applications, and network path dynamics. Different from these prior work, this paper leverages the availability of traffic flows exported from programmable home routers, and presents the first study of traffic characteristics of Internet-capable devices in home networks.

5 Conclusions and Future Work

In light of the rapid growth of home networks, this paper develops a traffic monitoring platform to collect and analyze network traffic for Internet-capable devices in home networks. Relying on programmable home routers that connect home networks to the Internet, we first collect network flow streams to traffic profiling servers. Subsequently, we analyze traffic characteristics of home networks, and use principal component analysis to uncover distinct clusters of application ports with temporal correlation. We are currently developing privacy-preserving data collection capacity into the traffic monitoring platform, so that we could deploy the platform into a large number of home networks to demonstrate its benefits in managing and securing home networks.

Acknowledgment. This research is supported in part by China 973 program under grant 2010CB328200, the Key Laboratory of Advanced Optical Communication Systems and Networks in Shanghai, and ASU New College SRCA grants.

References

1. CNN, Netflix takes up 32.7 of Internet bandwidth, http://www.cnn.com/2011/10/27/tech/web/netflix-internet-bandwith-mashable/index.html
2. Grinter, R., Edwards, K., Newman, M., Ducheneaut, N.: The Work to Make a Home Network Work. In: Proceedings of European Conference on Computer-Supported Cooperative Work (ECSCW) (September 2005)
3. Grinter, R., Edwards, K., Chetty, M., Poole, E., Sung, J., Yang, J., Crabtree, A., Tolmie, P., Rodden, T., Greenhalgh, C., Benford, S.: The ins and outs of home networking: the case for useful and usable domestic networking. ACM Transactions on Computer-Human Interaction 16(2) (2009)
4. Yang, J., Edwards, W.K., Haslem, D.: Eden: Supporting Home Network Management Through Interactive Visual Tools. In: Proceedings of ACM Symposium on User Interface Software and Technology (October 2010)
5. Poole, E., Edwards, K., Jarvis, L.: The Home Network as a Sociotechnical System: Understanding the Challenges of Remote Home Network Problem Diagnosis. Journal of Computer-Supported Cooperative Work Special Issue on CSCW, Technology, and Diagnostic Work (2009)
6. Aggarwal, B., Bhagwan, R., Das, T., Eswaran, S., Padmanabhan, V., Voelker, G.: NetPrints: Diagnosing Home Network Misconfigurations Using Shared Knowledge. In: Proceedings of USENIX Symposium on Networked System Design and Implementation (NSDI) (May 2009)
7. Cho, K., Fukuda, K., Esaki, H., Kato, A.: The Impact and Implications of the Growth in Residential User-to-User Traffic. In: Proceedings of ACM SIGCOMM (September 2006)
8. Dischinger, M., Haeberlen, A., Gummadi, K.P., Saroiu, S.: Characterizing Residential Broadband Networks. In: Proceedings of Internet Measurement Conference (October 2007)
9. Maier, G., Feldmann, A., Paxson, V., Allman, M.: On Dominant Characteristics of Residential Broadband Internet Traffic. In: Proceedings of Internet Measurement Conference (November 2009)
10. Xu, K., Wang, F., Lee, M.: HomeTPS: Uncovering What is Happening in Home Networks (Demo). In: Proceedings of IEEE Consumer Communications and Networking Conference (January 2012)
11. OpenWrt, OpenWrt: a Linux distribution for embedded devices, https://openwrt.org/
12. Lakhina, A., Crovella, M., Diot, C.: Diagnosing Network-Wide Traffic Anomalies. In: Proceedings of ACM SIGCOMM (2004)
13. Lakhina, A., Papagiannaki, K., Crovella, M., Diot, C., Kolaczyk, E., Taft, N.: Structural Analysis of Network Traffic Flows. In: Proceedings of ACM SIGMETRICS (June 2004)
14. Jolliffe, I.T.: Principal Component Analysis, 2nd edn. Spinger Series in Statistics (2002)
15. Edwards, W., Grinter, R., Mahajan, R., Wetherall, D.: Advancing the State of Home Networking. Communications of the ACM 54(6), 62–71 (2011)
16. Feamster, N.: Outsourcing Home Network Security. In: Proceedings of ACM SIGCOMM Workshop on Home Networks (HomeNets) (September 2010)
17. Yang, J., Edwards, W.K.: A Study on Network Management Tools of Householders. In: Proceedings of ACM SIGCOMM Workshop on Home Networks (HomeNets) (September 2010)

18. Calvert, K., Edwards, W.K., Feamster, N., Grinter, R.E., Deng, Y., Zhou, X.: Instrumenting Home Networks. In: Proceedings of ACM SIGCOMM Workshop on Home Networks (HomeNets) (September 2010)
19. Yiakoumis, Y., Yap, K., Katti, S., Parulkar, G., McKeown, N.: Slicing Home Networks. In: Proceedings of ACM SIGCOMM Workshop on Home Networking (August 2011)
20. DiCioccio, L., Teixeira, R., Rosenberg, C.: Impact of Home Networks on End-to-End Performance: Controlled Experiments. In: Proceedings of ACM SIGCOMM Workshop on Home Networks (September 2010)
21. Kreibich, C., Weaver, N., Nechaev, B., Paxson, V.: Netalyzr: Illuminating The Edge Network. In: Proceedings of Internet Measurement Conference (November 2010)
22. Chetty, M., Haslem, M., Baird, A., Ofoha, U., Sumner, B., Grinter, R.: Why Is My Internet Slow?: Making Network Speeds Visible. In: Proceedings of ACM Conference on Computer-Human Interaction (May 2011)
23. Hatonen, S., Nyrhinen, A., Eggert, L., Strowes, S., Sarolahti, P., Kojo, M.: An Experimental Study of Home Gateway Characteristics. In: Proceedings of ACM Internet Measurement Conference (November 2010)
24. Sundaresan, S., de Donato, W., Feamster, N., Teixeira, R., Crawford, S., Pescape, A.: Broadband Internet Performance: A View From the Gateway. In: Proceedings of ACM SIGCOMM (August 2011)